CAAStools, a toolbox to identify and test Convergent Amino Acid Substitutions.

*Fabio Barteri** [a,d]*, Alejandro Valenzuela** [a]*, Xavier Farré* [g]*, David de Juan* [a]*, Gerard Muntané* [a,e,f]*, Borja Esteve-Altava* [h] *and Arcadi Navarro* [a,b,c,d]*.*

Authors affiliations

- a. *IBE, Institute of Evolutionary Biology (UPF-CSIC), Department of Medicine and Life Sciences, Universitat Pompeu Fabra. PRBB, C. Doctor Aiguader N88, 08003 Barcelona, Spain*
- b. *Institució Catalana de Recerca i Estudis Avançats (ICREA) and Universitat Pompeu Fabra. Pg. Lluís Companys 23, 08010, Barcelona, Spain*
- c. *Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Av. Doctor Aiguader, N88, 08003 Barcelona, Spain*
- d. *BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, C. Wellington 30, 08005, Barcelona, Spain*
- e. *Hospital Universitari Institut Pere Mata, Institut d'Investigació Sanitària Pere Virgili (IISPV), Universitat Rovira i Virgili, Reus, Spain.*
- f. *Centro de investigación biomédica en red en salud mental (CIBERSAM), Spain.*
- g. *Genomes for Life-GCAT lab. Germans Trias i Pujol Research Institute (IGTP), Badalona, Spain*
- h. *European Molecular Biology Laboratory. Meyerhofstraße 1, 69117 Heidelberg, Germany*

** Equal contribution*

Abstract.

Background.

Upon phylogenetic data, coincidence of Convergent Amino Acid Substitutions (CAAS) with phenotypic convergences allow pinpointing genes that are likely to be associated with trait variation. Such findings can provide useful insights into the genetic architecture of complex phenotypes. Here we introduce CAAStools, a set of bioinformatics tools to identify and validate CAAS in orthologous protein alignments for pre-defined groups of species representing the phenotypic values targeted by the user.

Implementation and availability

CAAStools source code is available at http://github.com/linudz/caastools, along with documentation and examples.

Supplementary information

Supplementary data are available.

Introduction.

Convergent Amino Acid Substitutions (CAAS) can provide important insights into the genetic changes underlying phenotypic variation (Zhang and Kumar, 1997; Ray *et al*., 2015). Recent examples include the identification of genes potentially involved in marine adaptation in mammals (Foote *et al*., 2015), and the convergent evolution of mitochondrial genes in deep-sea fish species (Sheng *et al*., 2019). Notably, in 2018, Muntané *et al*. identified a set of 25 genes involved in longevity in primates (Muntané *et al*., 2018). A few years later, a similar analysis for a wider phylogeny retrieved 996 genes associated with lifespan determination in mammals (Farré *et al*., 2021). While these analyses often need to be adapted for each particular phenotype and phylogeny, the CAAS detection and validation strategies reported in the literature share some common steps (Rey *et al*., 2019). First, researchers select the species to compare for CAAS analysis, and split them into two or more groups according to the phenotype. These groups can be formed, for instance, by species having diverging values of a given continuous trait, or by species sharing different adaptations, like terrestrial and marine mammals (Foote *et al*., 2015). The second step consists in linking amino acid substitutions with each group. For that step, different approaches can be used, such as identifying identical substitutions for the same amino acid (Besnard *et al*., 2009; Chabrol *et al*., 2018), detecting topological incongruencies (Li *et al*., 2008), variations in amino acid profiles (Rodrigue *et al*., 2010; Rey *et al*., 2018), or relying on consistent patterns of groups of amino acids in different groups of species (Zhang et al, 2014; Muntané *et al*., 2018, Farré *et al*., 2021). The third step consists in testing the significance of the results. For instance, resampling tests can be performed to evaluate whether the number of detected CAAS is larger than expected by chance (Muntané *et al*., 2018; Farré *et al*., 2021), and also consistency checks with independent sets of species (Farré *et al*. 2021). Implementing a computational workflow for proteome-wide CAAS identification is challenging, as all these steps can become computationally expensive if applied to a large set of alignments. Here we present CAAStools, a toolbox for CAAS identification and validation, that is designed to be included in parallelized workflows.

Implementation.

CAAStools is a multi-modular python application organized into three tools. The outline of the suite is presented in **Figure 1**. The discovery tool is based on the protocol used in Muntané *et al*., 2018 and Farré *et al*., 2021. This approach identifies CAAS between two groups of species in an amino-acid Multiple Sequence Alignment (MSA) of orthologous proteins. These groups are named Foreground Group (FG) and Background Group (BG). Collectively, the two groups are called Discovery Groups (DG), as they represent the base for CAAS discovery. The CAAS identification algorithm scans each MSA and returns those positions that meet the following conditions: First, the FG and the BG species must share no amino acids in that position. Second, all the species in at least one of the two discovery groups (FG or BG) must share the same amino acid. The combination of these two conditions determines a set of different mutation patterns that the tool identifies as CAAS. Details on these patterns are provided in the Supplementary Table 1. Finally, CAAStools calculates the probability of obtaining a CAAS in a given position compared to randomized DGs, corresponding to the empirical p-value of the predicted CAAS in that position. The details of this calculation are presented in Supplementary section 3. The Resample tool sorts species into *n* virtual DGs (resamplings) for bootstrap analysis according to different permutation strategies. In a *Naive* modality, the probability of every species being included in a DG is considered identical and independent. However, species are phylogenetically related, biasing their probability of sharing a phenotype or amino acid. To address these phylogenetic dependencies CAAStools includes two other permutation strategies. In the *Phylogeny-restricted* modality, the randomization can be restricted to some taxonomic orders or defined clades (typically matching the phylogenetic divergences present within and/or between DGs). In the *Brownian motion* modality, resampling is based on Brownian Motion simulations. This *Brownian motion* modality is inspired by the “permulation” strategy for trait randomization (Saputra et al., 2021), and its implementation relies on the *simpervec()* function from the RERconverge package (Kowalczyk *et al*., 2019). Finally, the *bootstrap* tool determines the iterations returning a CAAS for each position in a MSA to establish the corresponding empirical p-value for the detection of a CAAS in that position. Both the discovery and the bootstrap tools are designed to be launched on single MSAs, in order to allow the user to parallelize the workflow for large protein sets.

Usage and testing

We tested CAAStools on the dataset from Farré *et al*., (2021). The details of this test are reported in Supplementary 3. The full dataset is available in the /test folder within the CAAStools repository.

Figures

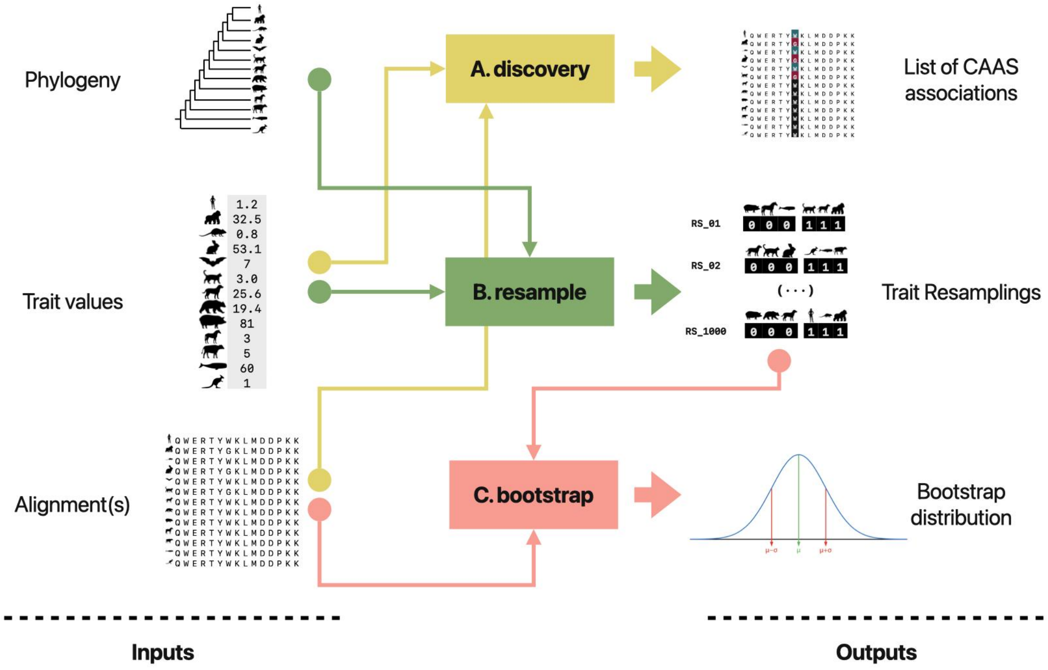

Figure 1. CAAStools layout. The 3 tools of the CAAStools suite rely on 3 pieces of information; a phylogenetic tree, the trait information, and an amino acid MSA. The discovery tool (A) detects the CAAS between two groups of species that are defined by the user on the basis of trait values. The resample tool (B) performs *n* trait resamplings in different modalities, on the bases of the phylogeny and the trait value distributions. The output of this resampling is processed by the bootstrap tool (C) that elaborates a bootstrap distribution from the MSA. All the tools can be executed independently.

Acknowledgements

We'd like to thank our collaborators at the Institute of Evolutionary Biology of Pompeu Fabra University for supporting us and sharing their ideas in scientific discussions, with particular mention of Dr. Tomas Marques Bonet and members of his group. This work was supported by I+D+i projects AEI-PGC2018-101927-BI00704 and PID2021-127792NB-I00 funded by MCIN/AEI/10.13039/501100011033 (FEDER Una manera de hacer Europa)" and by "Unidad de Excelencia María de Maeztu", funded by the AEI (CEX2018-000792-M) and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880).

Author Contributions

Fabio Barteri (FB) has been in charge of the development of CAAStools. Alejandro Valenzuela (AV) carried on the beta-testing and code debugging. FB and AV have contributed equally to this manuscript. Xavi Farré (XF), Gerard Muntané (GM) and Arcadi Navarro (AN) designed the CAAS identification and validation protocol. XF wrote a set of scripts to identify CAAS that served as template for CAAStools implementation. AN and GM conceptualized the method. AN, David de Juan (DJ) and Borja Esteve-Altava (BEA) - along with GM - participated in the scientific discussion and supervision of this project. DJ's contribution was particularly helpful for the optimization of CAAStools code. .

Bibliography

Besnard, Guillaume, A. Muthama Muasya, Flavien Russier, Eric H. Roalson, Nicolas Salamin, y Pascal-Antoine Christin. «Phylogenomics of C4 Photosynthesis in Sedges (Cyperaceae): Multiple Appearances and Genetic Convergence». Molecular Biology and Evolution 26, n.º 8 (1 de agosto de 2009): 1909-19. https://doi.org/10.1093/molbev/msp103.

Chabrol, Olivier, Manuela Royer-Carenzi, Pierre Pontarotti, y Gilles Didier. «Detecting the Molecular Basis of Phenotypic Convergence». Methods in Ecology and Evolution 9, n.º 11 (2018): 2170-80. https://doi.org/10.1111/2041-210X.13071.

Farré, Xavier, Ruben Molina, Fabio Barteri, Paul R H J Timmers, Peter K Joshi, Baldomero Oliva, Sandra Acosta, Borja Esteve-Altava, Arcadi Navarro, y Gerard Muntané. «Comparative Analysis of Mammal Genomes Unveils Key Genomic Variability for Human Life Span». Molecular Biology and Evolution 38, n.º 11 (1 de noviembre de 2021): 4948-61. https://doi.org/10.1093/molbev/msab219.

Foote, Andrew D., Yue Liu, Gregg W. C. Thomas, Tomáš Vinař, Jessica Alföldi, Jixin Deng, Shannon Dugan, et al. «Convergent Evolution of the Genomes of Marine Mammals». Nature Genetics 47, n.º 3 (marzo de 2015): 272-75. https://doi.org/10.1038/ng.3198.

Kowalczyk, Amanda, Wynn K Meyer, Raghavendran Partha, Weiguang Mao, Nathan L Clark, y Maria Chikina. «RERconverge: an R package for associating evolutionary rates with convergent traits». Bioinformatics 35, n.º 22 (1 de noviembre de 2019): 4815-17. https://doi.org/10.1093/bioinformatics/btz468.

Li, Gang, Jinhong Wang, Stephen J. Rossiter, Gareth Jones, James A. Cotton, y Shuyi Zhang. «The hearing gene Prestin reunites echolocating bats». Proceedings of the National Academy of Sciences 105, n.º 37 (16 de septiembre de 2008): 13959-64. https://doi.org/10.1073/pnas.0802097105.

Muntané, Gerard, Xavier Farré, Juan Antonio Rodríguez, Cinta Pegueroles, David A. Hughes, João Pedro de Magalhães, Toni Gabaldón, y Arcadi Navarro. «Biological Processes Modulating Longevity across Primates: A Phylogenetic Genome-Phenome Analysis». Molecular Biology and Evolution 35, n.º 8 (1 de agosto de 2018): 1990-2004. https://doi.org/10.1093/molbev/msy105.

Rey, Carine, Laurent Guéguen, Marie Sémon, y Bastien Boussau. «Accurate Detection of Convergent Amino-Acid Evolution with PCOC». Molecular Biology and Evolution 35, n.º 9 (1 de septiembre de 2018): 2296-2306. https://doi.org/10.1093/molbev/msy114.

Rey, Carine, Vincent Lanore, Philippe Veber, Laurent Guéguen, Nicolas Lartillot, Marie Sémon, y Bastien Boussau. «Detecting adaptive convergent amino acid evolution». Philosophical Transactions of the Royal Society B: Biological Sciences 374, n.º 1777 (22 de julio de 2019): 20180234. https://doi.org/10.1098/rstb.2018.0234.

Rodrigue, Nicolas, Hervé Philippe, y Nicolas Lartillot. «Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles». Proceedings of the National Academy of Sciences 107, n.º 10 (9 de marzo de 2010): 4629-34. https://doi.org/10.1073/pnas.0910915107.

Saputra, Elysia, Amanda Kowalczyk, Luisa Cusick, Nathan Clark, y Maria Chikina. «Phylogenetic Permulations: A Statistically Rigorous Approach to Measure Confidence in Associations in a Phylogenetic Context». Molecular Biology and Evolution 38, n.º 7 (1 de julio de 2021): 3004-21. https://doi.org/10.1093/molbev/msab068.

Shen, Xuejuan, Zhiqing Pu, Xiao Chen, Robert W. Murphy, y Yongyi Shen. «Convergent Evolution of Mitochondrial Genes in Deep-Sea Fishes». Frontiers in Genetics 10 (2019). https://www.frontiersin.org/articles/10.3389/fgene.2019.00925.

Zhang, Guojie, Cai Li, Qiye Li, Bo Li, Denis M. Larkin, Chul Lee, Jay F. Storz, et al. «Comparative genomics reveals insights into avian genome evolution and adaptation». Science 346, n.º 6215 (12 de diciembre de 2014): 1311-20. https://doi.org/10.1126/science.1251385.

Zhang, J., y S. Kumar. «Detection of Convergent and Parallel Evolution at the Amino Acid Sequence Level». Molecular Biology and Evolution 14, n.º 5 (mayo de 1997): 527-36. https://doi.org/10.1093/oxfordjournals.molbev.a025789.

CAAStools, a toolbox to identify and test Convergent Amino Acid Substitutions.

Supplementary information.

Supplementary 1. CAAS Discovery algorithm.

Given two Discovery Groups (DGs, Foreground and Background groups, FG and BG, respectively), the discovery tool recognizes as CAAS all those substitutions that meet two requirements. Let *A* be an MSA of *q* sequences of length *t*. We can describe A as an array of *t* positions [1]. Each position (*pos_i*) will consist of a set of N different amino acids, *a*, with absolute frequency (or count), *f*, where *NS* is the total number of symbols in the alignment [1].

[1] $$A = (pos_1\ pos_2\ \ldots\ pos_t\ )\ ;\ pos_i =\ a_1\ f_1\ a_2\ f_2\ \ldots\ a_{\ NS}\ \ldots\ f_{\ NS}\ \ ;\ \sum_1^{NS}\ \ f = q$$

The FG and the BG are formalized as sets of different species s_{FG} and s_{BG}, with no intersection and size l_{FG} and l_{BG} [2].

[2] $$s_{FG} = \left(s_1\ s_2\ \ldots\ s_{l_{FG}}\ \right)\ ;\ s_{BG} = \left(s_1\ s_2\ \ldots\ s_{l_{BG}}\ \right)$$

$$s_{FG} \cap s_{BG} = \emptyset$$

In each alignment position, s_{FG} and s_{BG} are associated with two sets of amino acids, $fg(pos_i)$ and $bg(pos_i)$, with length w_{FG} and w_{BG}.

[3] $$fg(pos_i) =\ a_1\ f_1\ a_2\ f_2\ \ldots\ a_{\ w_{FG}}\ \ldots\ f_{\ w_{FG}}\ \ \ ;\ bg(pos_i) = a_1\ f_1\ a_2\ f_2\ \ldots\ a_{\ w_{BG}}\ \ldots\ f_{\ w_{BG}}$$

CAAStools identifies a CAAS when three conditions are met [4]. First, the two groups must share no amino acids. This means that all the species in the FG need to have different AAs than the species in the BG. Second, at least one of the two DGs must share (or "converge to") the same amino acid. Also, the CAAS is detected if at least one amino acid is associated to both DGs

[4] $$CAAS_i\{fg(pos_i)\ \cap bg(pos_i) = \emptyset\ w_{FG}\ ==\ 1\ or\ w_{BG}\ \ ==\ 1\ w_{FG} > 0\ and\ w_{BG}\ >\ 0$$

The combination of these three rules defines 3 different mutation *patterns*. We define *pattern 1* when the DGs converge to two different amino acids ($w_{FG}\ =\ 1;\ w_{BG}\ \ =\ 1$). The *pattern 2* will be verified as the FG converges to one amino acid, but the BG will be associated with different amino acids ($w_{FG}\ =\ 1;\ w_{BG}\ >\ 1$). *Pattern 3* will consist in the opposite situation, or else when the FG is associated with different amino acids, whilst the BG converges to a single amino acid ($w_{FG} >\ 1;\ w_{BG}\ =\ 1$). **Supplementary Table**

1 summarizes the different mutation patterns and the meeting of requirements for CAAS identification.

Discovery Groups		Difference between DGs	Convergence in		Pattern
FG	**BG**		**FG**	**BG**	
KV	K	NO	NO	YES	Not a CAAS (No difference)
M	TM	NO	YES	NO	Not a CAAS (No difference)
MK	VE	YES	NO	NO	Not a CAAS (No convergence)
K	V	YES	YES	YES	**Pattern 1** (Both convergent)
K	VM	YES	YES	NO	**Pattern 2** (FG convergent, BG multiple)
KE	W	YES	NO	YES	**Pattern 3** (FG multiple, BG convergent)

Supplementary Table 1. *Mutation patterns and associated program decisions on CAAS assignment.*

Supplementary 2. CAAS discovery statistical testing

CAAStools calculates an empirical p-value for each CAAS prediction. This p-value is equal to the probability of obtaining a CAAS with random species, and under the same conditions as the CAAS discovery (size of the DGs, maximum permitted gaps and missing species). Following the MSA description in [1], we'll consider a couple of DG (FG and BG) of size l_{FG} and l_{BG}, as formalized in [2]. The probability to obtain a CAAS from random species is calculated as the probability of extracting concomitantly k_{FG} and k_{BG} objects from a population of size N over a number of extractions *n*, provided the conditions in [4], i.e. $k_{FG} \ \cap \ k_{BG} \ = \ \emptyset$ and $wk_{FG} \ == \ 1 \ or \ wk_{BG} \ == \ 1$ where *wk* is the number of symbols in the resampling *k*. This probability can be calculated through the probability mass function from the hypergeometric distribution [5].

[5] $$P(k) = \frac{\binom{K}{k}\binom{N-k}{n-k}}{\binom{N}{n}} = Hyp(N, K, k, n)$$

$$P(CAAS) = \ P(FG) \ * \ P(BG)$$

$$\{P(FG) \ = \ Hyp(N_{FG}, K_{FG}, k_{FG}, n_{FG}) \ N_{FG} = \ q \ - \ l_{BG} \ k_{FG} = l_{FG} \ - \ null_{FG} \ n_{FG} \ = \ k_{FG} \ ; \{P(BG) \ = \ Hyp(N_{BG}, K_{BG}, k_{BG}, n_{NG}) \ N_{BG} = \ q \ - \ l_{FG} \ k_{BG} = l_{BG} \ - \ null_{BG} \ n_{BG} \ = \ k_{BG}$$

Note that the size of the population N in each set is calculated as the difference between the total number of sequences in the alignment q and the size of the other group $(q \ - \ l_{BG}), (q \ - \ l_{FG})$, since the two extractions are concomitant but not interdependent. Also, the number of extractions k_{FG} and k_{BG} are equal to the difference between the size of the DGs and the number of indels and missing species allowed by the user (*null*). The terms K_{FG} and K_{BG} represent the number of successes in the population. In [6], [7] and [8], we see how this value can be calculated considering all the possible combinations of amino acid symbols that meet the requirements for CAAS detection [4].

[6] $$C_{P1,2} \ = \ \{K_{FG} = [f_j]; \ K_{BG} = [q - f_j] \ \forall \ a_j \ \in \ pos_i\}$$

$$C_{P1,2} \ = \ [(K_{FG_1}; \ K_{BG_1}), (K_{FG_2}; \ K_{BG_2}) \dots (K_{FG_z}; \ K_{BG_z})]$$

$$[7] \quad C_{P1,3} = \{K_{FG} = [q - f_j];\ K_{BG} = [f_j]\ \forall\ a_j \in pos_i\}$$
$$C_{P1,3} = [(K_{FG_1};\ K_{BG_1}), (K_{FG_2};\ K_{BG_2}) \dots (K_{FG_z};\ K_{BG_z})]$$

$$[8] \quad C_{P1} = \{K_{FG} = [f_j];\ K_{BG} = [f_h]\ \forall\ a_j, a_k \in pos_i\}$$
$$C_{P1} = [(K_{FG_1};\ K_{BG_1}), (K_{FG_2};\ K_{BG_2}) \dots (K_{FG_z};\ K_{BG_z})]$$

These combinations are based on patterns (*P*). Note that $C_{P1,2}$ and $C_{P1,3}$ overlap, and that the intersection coincides with C_{P1}. We can now calculate the CAAS probability separately for each pattern [9].

$$[9] \quad P(CAAS_{P1,3}) = \sum_{x=1}^{z} \quad Hyp(N_{FG}, K_{FG_x}, k_{FG}, n_{FG}) * Hyp(N_{BG}, K_{BG_x}, k_{BG}, n_{NG})$$
$$P(CAAS_{P1,2}) = \sum_{x=1}^{z} \quad Hyp(N_{FG}, K_{FG_x}, k_{FG}, n_{FG}) * Hyp(N_{BG}, K_{BG_x}, k_{BG}, n_{NG})$$
$$P(CAAS_{P1}) = \sum_{x=1}^{z} \quad Hyp(N_{FG}, K_{FG_x}, k_{FG}, n_{FG}) * Hyp(N_{BG}, K_{BG_x}, k_{BG}, n_{NG})$$

The probability to obtain a CAAS in position pos_i is hence calculated as it follows:

$$[10] \quad pvalue_{\ pos_i} = P(CAAS_{pos_i}) = P(CAAS_{P1,3}) + P(CAAS_{P1,2}) - P(CAAS_{P1})$$

Supplementary 3 – CAAS discovery from Farré et al., 2021.

As a test run for CAAStools, we repeated the CAAS discovery from the results published by Farré et al., in 2021 and entitled "*Comparative Analysis of Mammal Genomes Unveils Key Genomic Variability for Human Life Span*" (DOI: 10.1093/molbev/msab219). In this work, 13,035 MSA from UCSC public database (https://genome.ucsc.edu/, accessed August, 2019) were scanned to find CAAS between two groups of species with divergent maximum lifespan. The "long lived" group is formed by *Homo sapiens* (hg38), *Nomascus leucogenys* (nomLeu3), *Heterocephalus glaber* (hetGla2), *Myotis davidii* (myoDav1), *Myotis lucifugus* (myoLuc2), *Eptesicus fuscus* (eptFus1). The "short lived" group is formed by *Mesocricetus auratus* (mesAur1), *Rattus norvegicus* (rn6), *Pantholops hodgsonii* (panHod1), *Sorex araneus* (sorAra2), *Condylura cristata* (conCri1), *Monodelphis domestica* (monDom5). Farré et al., filtered the results from CAAS discovery to those CAAS having no gaps or missing species, and focused their analysis on the CAAS of scenarios 1 and 2, which correspond to patter 1 and 2 in CAAStools terminology.

We have repeated this analysis under the same conditions, filtering for pattern 1 and 2 and for no gaps in foreground (*short-lived* group) and background (*long-lived* group). The results (*Supplementary dataset 1*) and the phenotype configuration (*Supplementary dataset 2*) are available in the supplementary.material.xls spreadsheet. Our analysis confirmed the retrivement of 2737 mutations in 2004 MSA.